CROATICA CHEMICA ACTA
CCACAA **77** (1–2) 397–401 (2004)

ISSN-0011-1643
CCA-2941
Original Scientific Paper

Precise Annotation of Transmembrane Segments with Garlic – a Free Molecular Visualization Program*

Damir Zucić[a], and Davor Juretić[b]**

[a]*School of Medicine, University of Osijek, J. Huttlera 4, 31000 Osijek, Croatia*

[b]*Physics Department, Faculty of Natural Sciences, Mathematics and Education, University of Split, N. Tesle 12, 21000 Split, Croatia*

RECEIVED APRIL 8, 2003; REVISED NOVEMBER 14, 2003; ACCEPTED NOVEMBER 14, 2003

Comparison of transmembrane secondary structure prediction programs requires precise annotations of transmembrane segments of known proteins. Annotations that may be found in different databases differ significantly due to the lack of a standard method for annotation of transmembrane segments. Simple methods for the placement of a bilayer membrane with respect to the known 3D structure of membrane protein were developed to improve precise determination of the transmembrane portion of the protein. The methods were implemented in the freely available, open-source molecular visualization program called Garlic. The program, complete with the source code and extensive documentation, may be found at the site: http://garlic.mefos.hr/garlic.

Key words
membrane proteins
transmembrane part
molecular visualization
membrane model
hydrophobicity

INTRODUCTION

Precise comparison of different transmembrane secondary structure prediction methods[1–6] requires a number of proteins of known 3D structure for proper testing of each prediction method. In addition, it is necessary to define precisely the transmembrane segments of these proteins. However, even though the 3D structures are known in great detail for more than twenty unrelated membrane proteins, there are still significant differences in database annotations of transmembrane segments. There are several reasons for these differences.

There are significant thermal fluctuations of the molecules constituting the cellular membrane. The experiments with neutron and x-ray scattering[7] have shown that there are no sharp changes of charge density at any point across the membrane; thus there is no clear edge between the polar and apolar parts of the membrane. Most of the files with atomic models of phospholipid bilayers, widely available on the Internet, were created before the thermal fluctuations were experimentally measured and are not accurate.

Further, few molecular visualization tools[8–15] are capable of displaying any kind of a model of the cellular membrane attached to the protein structure. Most annotations of transmembrane segments are highly subjective. Some authors tend to designate the entire long alpha helices as transmembrane helices, while others take into account that the ends of such helices are certainly outside the membrane. There is also the problem of a relative orientation of the protein with respect to the mem-

* Dedicated to Professor Nenad Trinajstić on the occasion of his 65th birthday.

** Author to whom correspondence should be addressed. (E-mail: zucic@garlic.mefos.hr)

brane: the minimization of energy cannot be performed solely by visual inspection, without any calculation.

This article describes the method of identification of transmembrane segments, which was implemented in the Garlic program. Garlic is a free, open source molecular visualization program, designed in the first place for visualization and analysis of membrane proteins, though it may be used for visualization of any type of molecules, provided that data are available in the PDB format. Garlic is capable of automatically determining the orientation and placement of a given membrane protein with respect to the membrane. The membrane is represented as a pair of planes with adjustable separation. These planes represent the border between the hydrophobic core and the polar interface region. It is possible to define this border by using the time-averaged spatial distributions of the most important structural groups.[7] The transmembrane portion of the protein may be easily selected and manipulated independently.

METHODS

A portion of the cellular membrane is modeled as a pair of planes, represented in the figures by a pair of circles. The approximate thickness of the most hydrophobic part of the membrane is regarded as the separation of two planes. The orientation and placement of the membrane model with respect to the given membrane protein is performed using one of the two available methods.

The first method is suitable for small membrane proteins of the helix bundle type, such as rhodopsin, which lack two rings of aromatic residues at the edges of a hydrophobic region. The second method is suitable for large helix bundle proteins and for beta barrel proteins. Both types of proteins have a ring of aromatic residues at each edge of a hydrophobic region. Both methods have a very simple physical background, being essentially equivalent to energy minimization. However, the true molecular dynamics simulations will take much more time to execute, while the methods described here are relatively fast.

A typical helix bundle protein is assembled from a number of shorter chains. As the membrane insertion and oligomerization may be independent, it is not expected that any of the transmembrane helices will have a large hydrophobic moment. It is assumed that simple average hydrophobicity is just enough to recognize the transmembrane helices.

For small helix bundle proteins, the most hydrophobic portions of the structure are identified, using the chosen hydrophobicity scale. Three hydrophobicity scales are hard-coded in the program: Kyte-Doolittle scale,[16] Eisenberg consensus scale,[17] and Steven White octanol scale.[18] The most hydrophobic parts of the structure are used to prepare a set of unit vectors. These vectors are essentially the vectors that define the directions of their respective alpha helices. One of these vectors, associated with the most hydrophobic helix, is chosen as the initial unit vector, required to define the membrane. All other vectors are checked against this vector and multiplied by –1 if pointing to the opposite side. All vectors are combined into a single vector and the resultant is used as the improved membrane unit vector.

The geometric center of the whole structure is used as the initial position of the membrane center. An array of cells is associated with the symmetry axis of the membrane. All residues are projected to this axis, using the position of the CB atom as the representative point. For each cell, the total projected hydrophobicity is calculated. The average hydrophobicity is calculated for a number of adjacent cells. The width of a sliding window, multiplied by the width of a single cell, should be equal to the membrane thickness. The cell with the highest average hydrophobicity is taken to be the cell nearest to the membrane center.

In the next step, the membrane normal vector is refined. The initial normal vector is declined from its original direction to scan a set of alternative positions. For each direction, all residues are projected to the axis and the average projected hydrophobicity is calculated. Again, the sliding window width should be adjusted to cover the entire membrane thickness. The unit vector that gives the highest average hydrophobicity is taken as the new membrane normal vector. The refinement of the normal vector and membrane center is repeated once more.

The second method, designed for large helix bundle proteins and for beta barrel proteins, takes into account the fact that the inner sides of barrels might be quite polar. The same may be true of large helix bundle proteins, which may hide buried polar residues in the transmembrane region, close to the axis. Here again the geometric center is used as the initial membrane center. An arbitrary vector is used as the initial barrel symmetry axis. An auxiliary cylinder, divided into a number of equal patches, is associated with the axis. The side chains are projected to the surface of the auxiliary cylinder, searching for the most distant side chains. The side chains that point inside, with respect to the axis, are ignored, as these are not expected to be exposed to the membrane.

The cells are grouped into stripes, parallel to the cylinder axis. For each cell in each stripe, the average hydrophobicity is calculated over a given number of cells. The sliding window width, multiplied by the cell length, should be equal to the membrane thickness. The averaging process is repeated once more, this time grouping the cells into rings. Each ring lies in the plane perpendicular to the axis and covers the angle of 360°. The axis direction is varied, to cover the full space angle of 4π sterad, and the axis which gives the highest second average of hydrophobicity is taken to be the best one and is used as the membrane normal vector.

The membrane normal vector is refined searching for two rings of aromatic side chains (Phe, Trp and Tyr), which are typically found at the edges of transmembrane parts of beta barrel proteins. During this process, only the CG atoms of the aromatic side chains are projected. The side chains that point inwards with respect to the given axis are ignored. After the refinement of the membrane normal vector, the

membrane center is refined, projecting only aromatic side chains that point outwards.

Both methods were successfully tested with available, experimentally solved structures of both helix bundle and beta barrel types of proteins. Besides the atomic coordinates, the only parameter required to identify the transmembrane portions of proteins is the presumed thickness of the hydrophobic membrane core. For helix bundle proteins, a hard-coded default value of 3.0 nm (30 Å) may be used, while for beta barrel proteins the best value is 2.1 nm (21 Å). Two different values are used because most of the known beta barrel proteins belong to the outer membrane of Gram-negative bacteria, with thinner hydrophobic cores. The automatic attachment of the membrane model to the membrane protein takes only a couple of seconds on contemporary cheap computers, running Linux as the operating system.

The Protein Data Bank[19–20] (PDB) database already contains a large number of 3D structures of different membrane proteins, but the total number of unrelated proteins is relatively small. In April 2003, there were 16 unrelated sets of polytopic, large helical proteins and 7 unrelated sets of beta stranded proteins. Representative proteins from each group are listed in Table I.

The Garlic program may be ported to any unix or unix compatible system with the standard graphics interface, the X window system (X11R6 or X11R5). It is compliant with the POSIX industry standard and with the ANSI C language programming standard. It is highly portable across different hardware platforms. The program is also capable of drawing Ramachandran plots, helical wheel plots, Venn diagrams with statistics, averaged hydrophobicity and hydrophobic moment plots. It can be also used to compare two protein sequences, using a simple replacement matrix based on the Dayhoff PAM 250 matrix.[21] In addition to sixteen predefined coloring schemes, the residues may be colored according to the hydrophobicity or some other property. The latest version of the program, may be found at the site: http://garlic.mefos.hr/garlic-1,3/index.html, while the previous versions are already included in some of the most popular Linux distributions (Debian, LindowsOS, Linux MLD, Mandrake, and SuSE) as part of the standard software. The program, complete with the source code and extensive documentation, is released under GNU GPL (General Public License), which guarantees that all versions of this program and all its derivatives will remain free.

RESULTS AND DISCUSSION

The membrane placement methods implemented in Garlic work satisfactorily for most known membrane proteins. Figure 1 shows four different membrane proteins with attached membrane. A large gallery of color images may be found at the site: http://garlic.mefos.hr/gallery/tm/index.html, where most of the proteins listed in Table I may be found, complete with the membrane model. The method does not work properly for small or narrow pro-

TABLE I.

PROTEIN	SOURCE	PDB CODE
Photosynthetic reaction center	*Rhodopseudomonas viridis*	1PRC
Bacteriorhodopsin	*Halobacterium halobium*	1KME
Light-harvesting complex	*Rhodopseudomonas acidophila*	1KZU
Cytochrome c oxidase	bovine heart mitochondria	1OCC
Cytochrome bc1 complex	bovine heart mitochondria	1QRC
Fumarate reductase	*Escherichia coli*	1FUM
Potassium channel KcSA	*Streptomyces lividans*	1BL8
MscL channel	*Mycobacterium tuberculosis*	1MSL
MscS channel	*Escherichia coli*	1MXM
Bovine rhodopsin	bovine rod outer segments	1F88
Ca-ATPase	rabbit sarcoplasmic reticulum	1EUL
Aquaporin	human red blood cell	1FQY
Chloride channel	*Salmonella typhimurium*	1KPK
Formate dehydrogenase-N	*Escherichia coli*	1KQF
ABC transporter, BtuCD	*Escherichia coli*	1L7V
H^+-coupled efflux transporter	*Escherichia coli*	1IWG
8-stranded OmpA-fragment	*Escherichia coli*	1BXW
10-stranded protease OmpT	*Escherichia coli*	1I78
12-stranded phospholipase A	*Escherichia coli*	1QD5
14-stranded alpha-hemolysin	*Staphylococcus aureus*	7AHL
16-stranded porin	*Rhodobacter capsulatus*	2POR
18-stranded maltoporin	*Escherichia coli*	1MAL
22 stranded receptor FhuA	*Escherichia coli*	1FCP

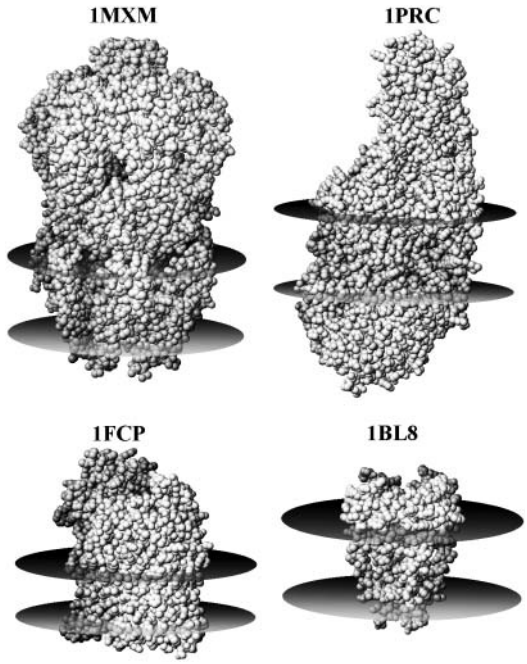

Figure 1. Four representative proteins: voltage gated mechanosensitive channel MscS (*E. coli*), photosynthetic reaction center (*Rhodopseudomonas viridis*), ferric hydroxamate uptake receptor (*E. coli*) and potassium channel (*Streptomyces lividans*).

teins, such as 1LOV and 1F88, which is expected for such a simple method. It is also possible that the orientation of narrow proteins is subject to significant fluctuations. For the vast majority of membrane proteins, it is possible to define precisely the transmembrane portions of the sequence and to use these data for training and testing procedures, performed during the development of the secondary structure predictions programs. The existence of a unique and simple method for annotation of transmembrane segments of known membrane proteins may improve the comparison of different transmembrane structure prediction programs. The method implemented in Garlic may be considered as a standard tool, because Garlic is the first modern molecular visualization program integrated into some of the most popular distributions of the Linux operating system and the first free and open program designed primarily for analyses of membrane proteins. At present, garlic is the only free visualization program capable of displaying and manipulating a simple geometric model for the biological membrane.

Acknowledgment.– Garlic development was supported by the Ministry of Science and Technology of the Republic of Croatia – Grant No. 0219031.

REFERENCES

1. G. von Heijne, *J. Mol. Biol.* **225** (1992) 487–494.
2. A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer, *J. Mol. Biol.* **305** (2001) 567–580.
3. T. Hirokawa, S. Boon-Chieng, and S. Mitaku, *Bioinformatics* **14** (1998) 378.
4. B. Rost, R. Casadio, P. Fariselli, and C. Sander, *Protein Science* **4** (1995) 521–533.
5. D. Juretić, D. Zucić, B. Lučić, and N. Trinajstić, *Computers Chem.* **22** (1998) 279–294.
6. D. Juretić, A. Jerončić, and D. Zucić, *Croat. Chem. Acta* **72** (1999) 975–997.
7. M. C. Wiener and S. H. White, *Biophys. J.* **59** (1991) 174–185.
8. C. W. V. Hogue, *Trends Biochem. Sci.* **22** (1997) 314–316.
9. R. A. Sayle and E. J. Milner-White, *Trends Biochem. Sci.* **20** (1995) 374–376.
10. D. C. Richardson and J. S. Richardson, *Protein Science* **1** (1992) 3–9.
11. N. Guex and M. C. Peitsch, *Electrophoresis* **18** (1997) 2714–2723.
12. R. Koradi, M. Billeter, and K. Wüthrich, *J. Mol. Graphics* **14** (1996) 51–55.
13. D. J. Bacon and W. F. Anderson, *J. Mol. Graphics* **6** (1988) 219–220.
14. E. A. Meritt and M. E. P. Murphy, *Acta Crystllogr. Sect. D* **50** (1994) 869–873.
15. W. Humphrey, A. Dalke, and K. Shulten, *J. Mol. Graphics* **14** (1996) 33–38.
16. J. Kyte and R. F. Doolittle, *J. Mol. Biol.* **157** (1982) 105–132.
17. D. Eisenberg, E. Schwarz, M. Komaromy, and R. Wall, *J. Mol. Biol.* **179** (1984) 125–142.
18. W. C. Wimley, T. P. Creamer, and S. White, *Biochemistry* **35** (1996) 5109–5124.
19. H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, and C. Zardecki, *Acta Crystallogr. Sect. D* **58** (2002) 899–907.
20. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *Nucleic Acids Res.* **28** (2000) 235–242.
21. M. O. Dayhoff, R. M. Schwarz, and B. C. Orcutt, *A Model of Evolutionary Change in Proteins*, in: M. O. Dayhoff (Ed.), *Atlas of Protein Sequence and Structure,* Vol. 5, National Biomedical Research Foundation, Washington, 1978, pp. 345–352.

SAŽETAK

Precizno označavanje transmembranskih segmenata s programom Garlic za vizualiziranje proteina

Damir Zucić i Davor Juretić

Usporedba računalnih programa za predviđanje transmembranske strukture membranskih proteina zahtijeva točno označavanje transmembranskih dijelova poznatih struktura proteina. Oznake koje se mogu naći u najpoznatijim bazama podataka značajno se razlikuju, zbog nedostatka standardne metode za označavanje transmembranskih dijelova proteina. U ovom radu su izložene jednostavne metode za postavljanje dvoslojne membrane u odnosu na molekulu membranskog proteina, u svrhu boljeg označavanja transmembranskih dijelova eksperimentalno riješenih struktura proteina. Te metode su ugrađene u Garlic – slobodni, otvoreni program za vizualizaciju proteina. Program, zajedno s detaljnom dokumentacijom, može se naći na Internet adresi: http://garlic.mefos.hr/garlic